A new data mining approach for the detection of bacterial promoters combining stochastic and combinatorial methods

Catherine Eng, Charu Asthana, Bertrand Aigle, Sébastien Hergalant, Jean-Francois Mari, Pierre Leblond

JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 16, Number 9, 2009

Pp. 1211–1225
DOI: 10.1089/cmb.2008.0122

Research Article

A New Data Mining Approach for the Detection of Bacterial Promoters Combining Stochastic and Combinatorial Methods

*CATHERINE ENG,[1,2] *CHARU ASTHANA,[1] BERTRAND AIGLE,[2] SÉBASTIEN HERGALANT,[1] JEAN-FRANÇOIS MARI,[1] and PIERRE LEBLOND[2]

ABSTRACT

We present a new data mining method based on stochastic analysis (Hidden Markov Model [HMM]) and combinatorial methods for discovering new transcriptional factors in bacterial genome sequences. Sigma factor binding sites (SFBSs) were described as patterns of *box1–spacer–box2* corresponding to the –35 and –10 DNA motifs of bacterial promoters. We used a high-order HMM in which the hidden process is a second-order HMM chain. Applied on the genome of the model bacterium *Streptomyces coelicolor* A3(2), the *a posteriori* state probabilities revealed local maxima or peaks whose distribution was enriched in the intergenic sequences (*"iPeaks"* for intergenic peaks). Short DNA sequences underlying the *iPeaks* were extracted and clustered by a hierarchical classification algorithm based on the SmithWaterman local similarity. Some selected motif consensuses were used as *box1* (–35 motif) in the search of a potential neighbouring *box2* (–10 motif) using a word enumeration algorithm. This new SFBS mining methodology applied on *Streptomyces coelicolor* was successful to retrieve already known SFBSs and to suggest new potential transcriptional factor binding sites (TFBSs). The well-defined SigR regulon (oxidative stress response) was also used as a test quorum to compare first- and second-order HMM. Our approach also allowed the preliminary detection of known SFBSs in *Bacillus subtilis*.

Key words: bacterial promoters, combinatorial methods, second-order hidden Markov models, stochastic model, *Streptomyces*, transcription factor binding site.

1. INTRODUCTION

THE VERSATILITY OF GENE EXPRESSION is essential for the adaptation of any living organism to its environment. A key level of control of gene expression is the first transcription step (i.e., initiation), promoted by interaction of RNA polymerase (RNAP) with gene promoter. RNAP holoenzyme is recruited at a given promoter through the recognition of a promoter by a transcriptional factor, called "sigma (σ) factor,"

[1]LORIA, UMR CNRS 7503 et INRIA Grand Est, Campus Scientifique, Vandœuvre-lès-Nancy, France.
[2]Laboratoire de Génétique et Microbiologie, UMR UHP-INRA 1128, IFR 110, Nancy Université, Faculté des Sciences et Techniques, Vandœuvre-lès-Nancy, France.
*These two authors contributed equally to this work.

which is a variable subunit of RNAP holoenzyme. Additional regulators can modulate the efficiency of this interaction and alter the transcriptional level.

The number of σ factors is variable in bacteria, and like other regulatory genes, it is related to the ecological niche occupied by the bacterium, and reflects its ability to cope with environmental changes including biotic competition. Beside the principal σ factor involved in the expression of the housekeeping genes, the other alternative σ factors are assumed to control sets of genes (called ''regulons'') involved in response to specific environmental stimuli. Most of the σ factors in bacteria are related to their homologue σ^{70} from *Escherichia coli*. In the free living soil bacterium *Streptomyces* model used in this study, not less than 60 σ factors are encoded and are involved in various responses leading to morphological as well as biochemical differentiation (Bentley et al., 2002; Ikeda et al., 2003). Furthermore, fewer than 10 σ factors binding sites have been experimentally characterized, including by determining their DNA binding sites (Bibb et al., 2000; Cho et al., 2001; Paget et al., 1998, 1999; Potuckova et al., 1995). The best characterized gene network, called the ''SigR regulon,'' is involved in oxidative stress response and includes about 30 genes (Paget et al., 2001).

The σ factors usually recognize two DNA boxes (about 6 bp long), constituting what is usually called the ''promoter.'' The bacterial promoters are located approximately at 10 and 35 bp upstream of the transcription start site. The spacer between these two boxes has a variable size from 16 to 20 bp for the major σ^{70} family. The sequence of a given promoter is typified by several levels of flexibility in addition to that of the spacer length. Mismatches can be tolerated and even allow for the modulation of promoter strength at some specific genes of the regulon. In some cases, an extended -10 promoter box may be observed and may substitute for the absence of a clear -35 element.

Experimental procedures are efficient to identify individual promoters but not conceivable for sets of genes at the whole genome scale. This motivated the search for computational methods based on the knowledge gained about the properties of known promoters or based on an efficient representation of DNA motifs by means of combinatorial or stochastic methods. There are three problems to solve when mining sigma factor binding sites (SFBSs) and generally transcriptional factor binding sites (TFBSs): (i) what kind of knowledge is available, (ii) how to locate a TFBS, and (iii) how to represent the variability of the extracted sequences. The use of prior knowledge about promoter sequences gives relevant data but can be extrapolated only to the same genome or to the more closely related strains or species. DNA stability can also be used as a characteristic of promoter regions (Kanhere and Bansal, 2005; Rangannan and Bansal, 2007).

The widely used approach for mining TFBSs is based on position weight matrices (PWM) trained on sets of nucleotide sequences known to be recognized by transcription factors (Hiard et al., 2007; Munch et al., 2005; Robison et al., 1998; Stormo, 2000). The limit of the method is the flexibility of the consensuses depending directly on the quantity of input biological data. In order to optimize the method, extra biological knowledge can be introduced into the parameters such as the UP-elements (A/T-rich regions upstream of the -35 box) (Typas and Hengge, 2005) considered in the Beadle method (Maetschke et al., 2006). Jacques et al. (2006) recently reported the use of matrices representing the genomic distribution of hexanucleotides pairs (*box1*-N_x–*box2*, where *box1* and *box2* ≥ 6) by estimating the ratio of hits in intergenic regions to the whole genome. Notably, the authors concluded that promoters are over-represented in intergenic regions. Other methods are trained with established descriptions of promoters such as Hidden Markov Model (HMM) (Jarmer et al., 2001; Petersen et al., 2003), support vector machine (SVM) (Gordon et al., 2006), or neural networks (Burden et al., 2005).

Statistical approaches are mostly based on the search for exceptional DNA motifs in subset of DNA sequences enriched or expected to be enriched in promoter sequences such as the upstream regions of co-regulated coding sequences (Bailey and Elkan, 1994; Buhler and Tompa, 2001; Lawrence et al., 1993; Van Helden et al., 2000; Segal and Sharan, 2005) or motifs significantly conserved in upstream regions of orthologous genes (Blanchette and Tompa, 2002; Loots et al., 2002; McCue et al., 2001; McGuire et al., 2000; Rajewsky et al., 2002; Siddharthan et al., 2005; Wang and Stormo, 2003). As an example, Sigffrid (Touzain et al., 2008) can find motifs significantly conserved in upstream regions of orthologous genes in different species. But the statistical methods cannot be used for organisms with insufficient biological knowledge or with raw genome sequences.

Other methods are based on an exact description of the promoters defined as structured motifs *(box1–spacer–box2)*. Vanet et al. (2000) specified an algorithm based on a suffix-tree to represent the input data and take them into account for the flexibility of the spacer by allowing jumps in the tree. This algorithm was improved by Carvalho et al. (2004) in terms of time and space but not sufficiently in terms of computation time. A similar approach is proposed by Eskin and Pevzner (2002), who first consider a

composite pattern (*box1–N_x–box2*) as one larger pattern (*box1* + *box2*) and then use mismatch trees to split pattern spaces in order to rule out weak subspaces. This method shows a good efficiency to identify long patterns. However, the composite patterns proposed by both approaches are rather long compared to the number of conserved motifs in the known promoters. Similar approaches were used by Studholme et al. (2004) and Mwangi and Siggia (2003), which compared the probabilities of all the patterns *box1–N_x–box2* with those of the motifs (*box1* and/or *box2*) upstream of coding sequences. Then, the composite patterns were clustered using sequence similarities to generate weight matrices (Li et al., 2002; Mwangi and Siggia, 2003; Studholme et al., 2004).

HMM has been used (Besemer et al., 2001; Churchill, 1989; Jarmer et al., 2001; Krogh et al., 1994; Liu et al., 2001; Nicolas et al., 2002; Thijs et al., 2001; Yada et al., 1998) for motif detection in two ways. The first approach is to assume that some information is encoded in the DNA sequence and is hidden by a background sequence that must be adequately modeled by an HMM (Liu et al., 2001; Thijs et al., 2001; Yada et al., 1998). These background models are capable of generating non-motif sequences from the genome under investigation. The second approach is to use HMMs for pattern matching or motif discovery. Here, a first-order HMM is developed as model codons and/or different patterns found in the genome (Besemer et al., 2001; Jarmer et al., 2001; Nicolas et al., 2002). The training procedure can be supervised (Jarmer et al., 2001) or unsupervised (Besemer et al., 2001), and the model can be targeted towards general motif search or carefully built for particular motifs like the SigA recognition site model (Jarmer et al., 2001).

Some review articles have carried out critical evaluations of motif discovery techniques and highlighted the limitations and potentials of the different methodologies (Osada et al., 2004; Tompa et al., 2005). In their review article, Hu et al. (2005) suggest that an ensemble algorithm approach, using analysis from different programs (methods), can complement one another's results and improve the prediction potential of DNA motifs.

We propose a new data mining method based on second-order HMM (HMM2) and combinational methods for SFBS prediction that voluntarily implements a minimum amount of knowledge. The original features of the presented methodology include (i) the use of k*mers* as observations, (ii) the Expectation Maximization estimation on the entire genome without *a priori* knowledge of their genetic content, (iii) an automatic peak extraction algorithm that captures the short DNA motifs underlying the peaks of the state *a posteriori* probability, and (iv) a suite of algorithms for finding SFBS and potential TFBS motifs.

On some points, our data mining method is similar to the work of Nicolas et al. (2002). Both methods use one HMM to model the entire genome. The parameter estimation is done in both cases by the EM algorithm. Both methods look for attributing biological characteristics to the states by analyzing the state output *a posteriori* probability. But our method differs on the following points: we use (i) an HMM2 that has proven interesting capabilities in modelling short sequences, and (ii) an original peak extraction algorithm to locate some short nucleotides sequences that could be a part of TFBS motifs (*box1* or *box2*). These sequences are further reassembled in TFBS composite motifs of the form *box1–spacer–box2* by means of combinatorial methods.

2. METHODS AND ALGORITHMS

2.1. HMM2 specifications

HMM2 was introduced by He (1988). As shown in speech recognition, modeling of the hidden process by a second-order Markov chain exhibited a good capability in representing short segments (Du Preez, 1998; Mari et al., 1997). In Data Mining (Mari and Le Ber, 2006) and Ecology (Le Ber et al., 2006), better performances were achieved by considering the contextual information defined by the neighboring observations. In our case (genomics), the nucleotides at index $t-1, \ldots, t-k+1$ define the contextual information that leads to the definition of k*mers*. The higher is the k value, the more important is the contextual information. It should be noted that k is not the order of the hidden Markov chain.

2.1.1. HMM2 mathematical definitions.

An HMM2 is defined by:

- a set S of N states, $\boldsymbol{S} = \{s_1, s_2, \ldots, s_N\}$;
- a transition matrix $\boldsymbol{A} = (a_{i_1 i_2 i_3})$ over S×S×S where $a_{i_1 i_2 i_3}$ is the *a priori* transition probability $P(X_t = s_{i_3} | X_{t-2} = s_{i_1}, X_{t-1} = s_{i_2})$ for the hidden process to be in state s_{i_3} at index t assuming it was in the state s_{i_2} at index $t-1$ and s_{i_1} at index $t-2$;

- a matrix $\boldsymbol{B}$ that represents the N discrete probability functions (*pdf*) over the set of M output symbols (the k*mer*). $\boldsymbol{B}(i,o) = P(O_t = o | X_t = s_i)$. $\boldsymbol{B}(i,o)$ is the conditional probability of symbol (k*mer*) o assuming the state s_i.

An HMM2 model is defined by a second-order Markov chain X that governs a set of *pdf* of output symbols. We have investigated the use of k*mer* as output symbols instead of nucleotides. A k*mer* may be viewed as a single nucleotide y_t observed at index t with a specific context $y_{t-k+1}, \ldots, y_{t-2}, y_{t-1}$ made of $k-1$ nucleotides that have been observed at index $t-k+1, \ldots, t-1$. Similarly, a DNA sequence can be viewed as a sequence of overlapping k*mer* that an HMM analyzes with a consecutive shift of ℓ. For example, a sequence *##TAGGCTA* can be viewed as a sequence having the same length of 3-*mer* ($k = 3$ and $\ell = 1$): ##T-#*TA-TAG-AGG-GGC-GCT-CTA*, where # represents an empty context.

The HMM2 training is performed by the EM algorithm. The EM algorithm is an efficient iterative procedure (Dempster et al., 1977) that locally maximizes the Maximum Likelihood (ML) estimate of P(O = o | *HMM2*). Unfortunately, the maximum value depends on the initial conditions. The HMM2 estimation formulas implemented by the EM algorithm (Baum et al., 1970) generalize the classical formulas that are used to estimate a first-order HMM. Basically, given a sequence of symbols $o_1, o_2, \ldots, o_T$, the second-order EM algorithm performs the expected count of the transition $s_{i_1}, s_{i_2}, s_{i_3}$ at index $t-2$, $t-1$, t:

$$\eta_t(i_1, i_2, i_3) = \mathrm{P}(\mathrm{X}_{t-2} = s_{i_1}, \mathrm{X}_{t-1} = s_{i_2}, \mathrm{X}_t = s_{i_3} \big| O_1^T = o_1^T). \tag{1}$$

The re-estimated second-order *a priori* transition probabilities are normalised as follows:

$$\overline{a_{i_1 i_2 i_3}} = \frac{\sum_t \eta_t(i_1, i_2, i_3)}{\sum_{i,t} \eta_t(i_1, i_2, i)}. \tag{2}$$

Whereas the re-estimated first-order *a priori* transition probabilities are normalised as follows:

$$\overline{a_{i_2 i_3}} = \frac{\sum_{i_1,t} \eta_t(i_1, i_2, i_3)}{\sum_{i_1,i,t} \eta_t(i_1, i_2, i)}. \tag{3}$$

The re-estimated output probability $\boldsymbol{B}(i,o) = P(O_t = o | X_t = s_i)$ is computed as in a HMM1:

$$\overline{B(i,o)} = \frac{\sum_{i_1,i_2,t} \eta_t(i_1, i_2, i) 1_{o=o_t}}{\sum_{i_1,i_2,t} \eta_t(i_1, i_2, i)} \tag{4}$$

where $1_{o=o_t}$ means 1 if the condition $o = o_t$ is true, zero otherwise. When o is a k*mer*, $o = w_1 w_2 \ldots w_k$ of k nucleotides $w_1, \ldots, w_k$, additional normalization constraints are usually (Bize et al., 1999) introduced to get the output conditional probability $P(w_k | w_1, \ldots, w_{k-1}, X_t = s_i)$ that expresses the dependencies between the k nucleotides on state $\mathrm{s_i}$. This probability is re-estimated as follows:

$$\overline{B(i, w_k, \ldots, w_1)} = \frac{\sum_{i_1,i_2,t} \eta_t(i_1, i_2, i) 1_{y_t = w_k, y_{t-1} = w_{k-1}, \ldots, y_{t-k+1} = w_1}}{\sum_{i_1,i_2,t} \eta_t(i_1, i_2, i) 1_{y_{t-1} = w_{k-1}, \ldots, y_{t-k+1} = w_1}}. \tag{5}$$

We have observed (data not shown) that these constraints dramatically smooth the state *a posteriori* probability $P(X_t = s_i | Y_1^T = y_1^T)$. On the other hand, when these constraints are not implemented, the model cannot generate coherent sequences except if the k*mer* are not overlapping. Because we do not use HMM to generate sequences of nucleotides, we do not implement these constraints.

Following Nicolas et al. (2002), we used a single ergodic HMM2—all the states are connected together—to model the entire genome, and estimated its parameters using Eqs. (2) and (4) of the second-order EM algorithm. During the last iteration of the second-order EM algorithm, we performed the *a posteriori* decoding. At each index t (at each k*mer* position in the DNA sequence), we computed the *a posteriori* probability $P(X_t = s_i, X_{t-1} = s_i |$ *"the sequence"*) for a specific state s_i, and look for the fluctuations of this probability along the genome. The sudden increase in this *a posteriori* probability locates a short DNA sequence that is typical of state s_i (called "heterogeneity"). Obviously, to extract such heterogeneity, the *pdf* associated with the hidden states must be different. We used the Kullback-Leibler (1951) information number—called "divergence"—between two distributions as a distance between two

states. We have designed a training strategy by successively training an ergodic HMM2 with an increasing number of states until a state appears to be distant enough from the other states. This state—called "best decoding state"—will be used further to detect local heterogeneities. The other *a posteriori* probabilities associated to the nondecoding states are disregarded.

2.1.2. HMM2 biological input preprocessing. In contrast to the pattern matching paradigm, in which the HMMs are used to discriminate between various and well-identified motifs, only two models were specified in our experiments. These models do not incorporate any *a priori* knowledge of the genetic structure of the genome of interest. The complete bacterial genome is used to train (maximum likelihood estimation using the second-order EM algorithm) two specific HMM2: HMM2+ and HMM2−. Ideally, when a marked GC skew is observed, as in the case of *B. subtilis* (Kunst et al., 1997), the HMM2+/− models were constructed to incorporate the biased base composition of DNA strands relative to the position of the replication origin. In contrast, when the genome does not show a definite GC skew, as in the case of *S. coelicolor* (Bentley et al., 2002), the HMM2+ and HMM2− models were constructed corresponding to the 5′ to 3′ sequence of the linear chromosome and its reverse complement, respectively. The best decoding state is identified for both HMM2+ and HMM2− models.

2.1.3. Automatic peaks detection and underlying sequence extraction. A peak is detected as local maximum of the *a posteriori* probability values generated over the genome, characterized by its width and its height (Fig. 1). The search of *a posteriori* probability peaks for a given state is performed using a sliding window 200-nucleotide-long window with an overlap of 100 nucleotides. The height of a peak is defined by its maximum value, which must be higher than the mean plus a given percentage of the dynamics (maximum–minimum values). This percentage is called "*peak-variation.*" Once a peak is located, the short DNA sequence underlying the peak (called "*Peak-motifs*") is automatically extracted. Both HMM2+ and HMM2− models are processed to get a set of total *Peak-motifs*. Peaks in the intergenic regions are called "*iPeaks*," and the underlying short DNA sequence is called a "*iPeak-motif.*"

2.1.4. Clustering the iPeak-motifs. In order to detect similar *iPeak-motifs,* we used an unsupervised clustering algorithm strategy (the ClusterMean algorithm; Algorithm 1), with the mean distance hierarchical grouping algorithm described by Ward (1963). This procedure builds a hierarchy where the closest sequences are in the same cluster. The distance between two sequences is defined as the inverse of their similarity, and the distance between two clusters C_1 and C_2 is defined as the mean distance between the pairs $(sequence_{i1}, sequence_{i2})$, where $sequence_{i1}$ belongs to C_1 and $sequence_{i2}$ belongs to C_2. The algorithm builds a hierarchy of distinct (non-overlapping) clusters that are represented by a consensus motif. Each cluster consensus motif was generated by Multalin (Corpet, 1988), and its significance was statistically validated by R'MES, a tool for finding exceptional motifs in sequence (Hoebeke and Schbath, 2006). R'MES score is a statistical score of exceptionality, where frequent motifs have higher scores.

Algorithm 1 The ClusterMean algorithm

1. Make a N square distance matrix (N = all *iPeak*-motifs). The distance, d = 1/s, where s is the Smith and Waterman (1981) score for the local alignment between a given pair of *iPeaks*.
2. Cluster *iPeaks* into clusters using the mean distance hierarchical grouping, based on their distance values between clusters.
3. Find the consensus sequence representing each cluster of motifs using the programme Multalin.
4. Check for redundancy (group clusters with same consensus). Select statistically significant cluster consensus using the programme R'MES to verify whether a consensus is a good representation of its cluster.

2.2. SFBS consensus search algorithm

A SFBS is structured as *box1-spacer-box2* where the *spacer* ranges from 3 to 25 (Algorithm 2). In order to retrieve SFBSs, the basic idea of the mining strategy, is to select a cluster having a high R'MES score consensus, extend all the sequences belonging to this cluster and look for over-represented motifs using R'MES. The consensus of the cluster acts for *box1*, the shorter motifs spaced with appropriate spacer value(s) act for *box2*.

Algorithm 2 Modeling SFBS motif

```
foreach box1 (consensus motif of a cluster) do // Step1: Fixing box1
    extend all the sequences in the cluster to length of 50;
    use R'MES to find statistically significant over-represented boxes of length,
    l_box2 (these boxes act as potential box2);
      foreach1 box2 given by R'MES do // Step2: Finding box2
            foreach2 spacer value in a given range do
              scan_for_match (Dsouza et al., 1997) "box1-spacer-box2" in the whole genome
              compute the number of occurrences
              if the number of occurrences ≥threshold then
                    log on a file "box1-spacer-box2", the number of occurrences found,
                    the value of spacer
                    and the R'MES score.
                    (e.g. GAAT-spacer-GGT, number of occurrences = 2, spacer = 6,8).
              endif
            end foreach2
      end foreach1
    Rank the TFBSs based on their number of occurrences and their R'MES score in descending order
    Select the N first
    end foreach
Compile together the files, this gives a list of putative SFBSs and other TFBSs (than SFBS) that can be probed
for further investigation
Define a class with the genes that are found downstream each TFBS by scanning the whole genome (or intergenic
region) with the pattern matching program: scan_for_match.
```

2.3. General parameters selection using the set-sco

In order to tune (i) the HMM2 parameters (order, topology and number of iterations of the EM algorithm), (ii) the peak extraction algorithm parameters, (iii) the number of classes during the clustering process, and (iv) the thresholds used by R'MES to detect over-represented motifs, we specified a manageable validation dataset. For this purpose, the *S. coelicolor* A3(2) genome (8.7 Mb) was fragmented into 174 non-overlapping segments of about 50 kilobases. We defined our validation data (called "*set-sco*") as the minimal set of fragments that includes all the genes regulated by the sigma factor SigR involved in oxidative stress response (Paget et al., 2001) scattered along the genome. SigR regulon includes at least 30 genes and is the largest regulon experimentally validated in *Streptomyces*. The *set-sco* covered 1.15 MB and contained 1135 genes, including the 30 SigR-regulated genes. Out of these 30 SigR genes, 25 have their −35 box starting in an intergenic DNA sequence (Table 1), while the remaining five genes have −35 box in the upstream coding sequences (Paget et al., 2001).

2.3.1. HMM order. Table 2 shows the specificity and the sensitivity at binding sites level by specifying the following values: TP (true positives), the number of binding sites predicted as binding sites; TN (true negatives), the number of nonbinding site predicted as nonbinding sites; FP (false positives), the number of nonbinding sites predicted as binding sites; and FN (false negatives), the number of binding sites predicted as nonbinding sites.

The specificity and the sensitivity of the SigR binding sites (−35 box) recognition using the *set-sco* are 0.24 and 0.8, respectively, with the HMM2 analysis. Using a HMM1 model, only 17 over 25 entire known sigma factor SigR binding sites (−35 box) localized in intergenic regions are detected, in contrast to the HMM2 model where all −35 boxes have been found (+/− 2nt; Table 1). The specificity of HMM1 (0.23) and HMM2 (0.24) are comparable. However, the HMM2 model makes less false negatives errors and shows a better sensitivity (0.8) compared to the HMM1 (0.68).

2.3.2. HMM topology. We found in *set-sco* analysis that four states and 3-*mers* HMM2 model was the best topology. The divergence matrix identified two different decoding states for HMM2+ and HMM2− models. The number of training iterations was fixed to five and gave the best enrichment in

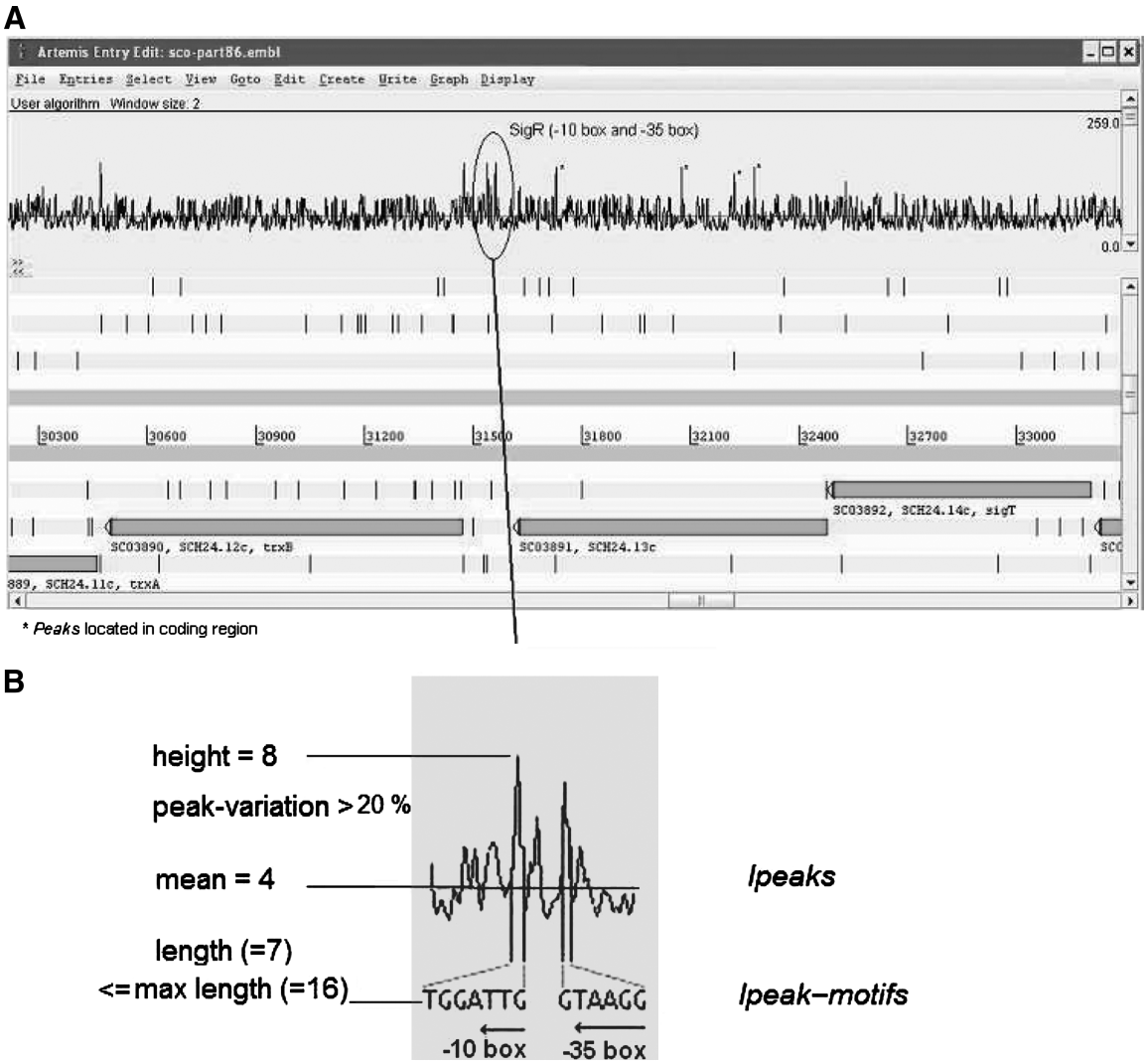

FIG. 1. Visualization of the HMM2 *a posteriori* output of *S. coelicolor.* (**A**) Figure visualized with Artemis (Rutherford et al., 2000). The top graph shows HMM2 output; the bottom shows the annotated physical sequence (using the EMBL file). (**B**) Zoom of SCO3890 *iPeaks.* The *iPeak-motifs* show the sigma factor SigR −35 box (GGAAT) and −10 box (GTT) motifs. Peak characteristics (*peak-variation* and width) are marked in the figure.

intergenic regions relative to the coding sequences (3.8-fold; for additional data, see *set-sco_analysis*; see online Supplementary Material at *www.liebertonline.com*).

2.3.3. Peak extraction parameters. The parameters of extraction of the *iPeak-motifs* were selected to find the maximum number of SigR binding sites. Better detection of the SigR −35 box motif was obtained with a value of *peak-variation* set to 20% (Fig. 1). Table 1 shows the *iPeak-motifs* detected for *set-sco* with *peak-variation* set to 20%. All −35 boxes of the SigR binding sites are detected. The maximum width of the *iPeak-motifs* was fixed to 16 nucleotides. Less than 0.6% of the peaks were discarded by applying this threshold. Discarding these peaks noticeably improved the clustering process.

2.3.4. Clustering iPeak-motifs. Different similarity measures (local and global) were tested in the clustering. The Smith and Waterman local similarity (1981) gave the more homogeneous clusters. We empirically observed on the *set-sco* that the optimum number of sequences in a cluster was roughly 40–50. Deviation from this optimal cluster size skewed the consensus determination. With larger sizes, no clear consensus could be determined while in smaller clusters, the same consensus would appear in several

Table 1. Detection of SigR Binding Sites (−35 Box: GGAAT) Using HMM1 and HMM2

Gene_name	*Strand*	*Intergenic*	*HMM2*	*HMM1*
SCM1.15	1	✓	AGT**GGAA**(t)	X
Trx A3, SCM1.18	1	✓	GCGC**GGAAT**AC	CGC**GGAAT**ACC
SC6D7.18c, RbpA	1	✓	CG**GGAAT**CTTT	X
RelA, SCL2.03	−1	✓	CGGA**GGAAT**	A**GGAAT**C
7H2.09c	−1	✓	ACAC**GGAAT**AG	AC**GGAAT**AGCG
7H2.11c	−1	✓	TCCCG**GGAAT**GCC	CCG**GGAAT**GCC
6G10.34c	−1	✓	TCCC**GGAAT**GAAT	CC**GGAAT**GAATC
8E4A.04c	−1	✓	GCCG**GGAAT**GG	CCG**GGAAT**GGG
PepN	1	—	—	—
E20.23	1	✓	CCGG**GGAAT**	X
E19A.11c	−1	—	—	—
ARNt 3226540	−1	✓	CGCCG**GGAAT**AGG	CCG**GGAAT**AGGCT
E25.24c	−1	✓	TGAGG**GGAAT**C	G**GGAAT**C
E87.13	1	✓	TGA**GGAA**(t)	X
E22.04	1	✓	CCGC**GGAAT**AG	CCGC**GGAAT**AGGTC
HrdD	−1	✓	GTTG**GGAATT**C	TTG**GGAATT**CT
FoIE	−1	✓	GCCC**GGAAT**GT	CC**GGAAT**GT
E9.22	1	✓	CTG**GGAA**(t)	T**GGAAT**A
TrxB, H24.12c	−1	✓	GCG**GGAAT**G	CG**GGAAT**GC
GuaB	1	✓	GT**GGA**(at)	X
P8.26c	−1	✓	AGCGTAG**GGAATGTT**	**GGAAT**
MoeB	1	✓	TCTC**GGAAT**GAAAAAG	X
23B6.11c	−1	—	—	—
SigR	1	✓	GCCTG**GGAAT**G	X
RpmE3, 6G5.03	1	—	—	—
3D11.22	1	✓	GCGGC**GGAAT**AGC	GC**GGAAT**AGCC
CinA	1	—	—	—
HflX	1	✓	CCG**GGAA**(t)	CCCCG**GGAAT**CTC
4B501c	−1	✓	TCA**GGAAT**G	GTCA**GGAAT**GCGTC
6A5.08	1	✓	CTCA**GGAAT**	X

—, not available, the −35 box of SigR binding site is located in coding sequence; X, not found.

clusters. During the Ward algorithm step, we used SigR SFBS as a marker. The process of merging two clusters, whose consensus motif contained **GGAAT,** is stopped when the motif consensus after merging does not contain **GGAAT**. The average size of the clusters was 50 sequences, and **GGAAT** still appeared in five clusters. In these five clusters, we observed that the motif **GGAAT** has a R'MES score higher than six, so this value was retained for assessing the exceptionality of five nucleotide motifs in further experiments. For a five letters motif, the R'MES threshold is 4.42 with 99.5% confidence (i.e., all motifs having score higher than this threshold are exceptional). This method assesses the validity of the consensus found by Multalin. These values were used in all further experiments, and represented the default parameters of the data mining. Extrapolating on the entire genome, the 16,913 *iPeak* DNA motifs were grouped into 360 clusters (i.e., approximately 47 sequences per cluster).

Table 2. Comparison of the SigR Binding Sites (−35 box = GGAAT) Recognition between HMM1 and HMM2 on the *set-sco*

	HMM2	*HMM1*
Sensibility[a]	0.8	0.68
Specificity[b]	0.24	0.23

[a]Sensitivity = TP/(TP+FN).
[b]Specificity = TN/(FP+TN).
TP, true positive; FN, false negative; FP, false positive; TN, true negative.

3. RESULTS AND DISCUSSION

Genomes of bacteria were obtained from the EMBL (Emmert et al., 1994) and were used to extract the raw sequence data used in this study as well as their corresponding annotation. DBTBS (Makita et al., 2004) was used for *B. subtilis* regulon analysis. The HMM2 core library (CarottAge: *www.loria.fr/~jfmari/App/*) was developed in C++, and included all the algorithms for modelling genomic sequences used in this study. A graphical interface was also developed in C++/XWindow to create, configure and display the model files, set the k*mer* reading methods, dump all the HMM physical topologies and screen their selected *a posteriori* probabilities alongside the sequences. The clustering software was developed in Java, and data mining and data formatting were done using Perl scripts. For additional results and data, see online Supplementary Material at *www.liebertonline.com.* All algorithms run on a P4 equivalent desktop computer, provided that the various parameters are set to reasonable values.

3.1. Identification of the dagA gene's promoters

The peak extraction algorithm was tested on the *dagA* gene. This gene was chosen, as it is not included in *set-sco* and it is not regulated by SigR. The *dagA* gene encodes the extracellular agarase precursor and is known to be controlled by four different promoters recognized by at least three (and probably four) different RNA polymerase holoenzymes (Buttner et al., 1988; Servin-Gonzalez et al., 1994). Using the HMM2 trained on the whole genome and the peak extraction algorithm—both tuned on *set-sco*—10 motifs could be extracted. Three of the previously known promoters were fully detected by our method. This includes *dagA*-p2 (ATTGTCA-N16-GTAGCATTC), *dagA*-p3 (GGAACTTT-N15-CTCTCGAAT), and *dagA*-p4 (TATAAGA); the motifs in brackets correspond to the previously identified binding sites. For *dagA*-p1 (TGGAGC-N18-TGGAATGA), only the −10 box (TGGAATGA) was found (for additional data, see *dagA analysis*; see online Supplementary Material at *www.liebertonline.com*). These results confirm the validity of the parameters tuned on the *set-sco*.

3.2. Selecting infrequent occurrences

Searching for under-represented TFBS consensuses can be done by selecting the R'MES score threshold to select infrequent occurrences (higher negative scores). The extracted *iPeak-motifs* exhibited the BldN binding consensus and identified the unique gene, annotated *bldM* (SCO4768), reported as a BldN target (Bibb et al., 2000).

3.3. Annotating SFBSs from S. coelicolor genome

This data mining method was next applied to the whole genome of *S. coelicolor*. We extracted 16,913 *iPeak* DNA motifs grouped into 360 clusters. Each cluster had a significant motif sequence according to R'MES, and after filtering duplicated consensus, 357 motifs were obtained. So, 228 out of 357 motifs holding out R'MES score higher than six (see Section 2.3), were used for further analysis. Table 3 shows a partial list of cluster consensus motifs that contain a box part (−35 or −10) of already known SFBSs like SigB binding sites and WhiG binding sites.

The parameters *spacer* and l_{box2} were restricted to a range of values consistent with biological knowledge of the SFBSs. For all the 228 motifs, the number of possible consensuses is the product between the *spacer* and *length* of box2, which equals 8,280 possibilities for $3 \leq spacer \leq 25$, and $3 \leq l_{box2} \leq 5$. To reduce computation time from a couple of days to a few hours, the range of *spacer* and of the R'MES threshold was restricted ($14 \leq spacer \leq 20$). We found 6,236 *box1-spacer-box2* putative consensuses. Finally, applying the modeling SFBS motif algorithm (see Section 2.2), 812 potential TFBS consensuses were suggested and are available, including known SFBS consensuses (for additional data, see online Supplementary Material at *www.liebertonline.com*).

Table 4 shows the main results of 812 predicted TFBSs:

- Five motifs that include or overlap the SigR consensus (GGAAT-N_{18}-GTT) were found.
- Three putative WhiG binding sites (Chater et al., 1989) were found. The extracted set of genes found downstream of these consensuses includes the two previously biologically determined targets of *whiG, whiI* (SCO6029), and *whiH* (SCO5819) (Ainsa et al., 1999; Ryding et al., 1998).

Table 3. Partial List of Cluster Motifs Defined from ClusterMean Algorithm

No.	*TFBS box*	*Known motif*	*Predicted motif*	*R'MES score*
1. SigB	−35	ANGNNT	AAGATt	7,3824
1. SigB	−35	ANGNNT	ACGGCTt	8,4372
1. SigB	−35	ANGNNT	AGGCTt	8,4699
1. SigB	−35	ANGNNT	AGGATt	8,3462
1. SigB	−10	GGGTA	CGGGTa	8,2924
2. SigE	−10	TCTY	GCTCTt	6,628
2. SigE	−10	TCTY	ATCTT	7,9511
3. SigR	−35	GGAAT	CGGGAat	8,2738
3. SigR	−35	GGAAT	TCGGAAt	8,273
3. SigR	−35	GGAAT	GGGGAat	8,026
3. SigR	−35	GGAAT	GCGGAAt	8,2233
3. SigR	−35	GGAAT	AGGGAAt	8,2155
3. SigR	−10	GTT	GCGTTA	8,286
3. SigR	−10	GTT	aCCGTTt	8,2254
3. SigR	−10	GTT	CTGTTT	8,3584
4. WhiG	−35	TRVR	cCTGAAa	8,1572
4. WhiG	−35	TRVR	GCTGAa	8,201
4. WhiG	−35	TRVR	TGGATt	8,2126
4. WhiG	−35	TRVR	cTTGAAt	7,2117
4. WhiG	−35	TRVR	TGCGAA	8,2401
5. PTF1	−35	GAAC	GAACTt	6,332
5. PTF1	−10	GTTG	gTTGAa	7,901
6. PTF2	−35	TGGT	cTGGTAa	6,062
6. PTF2	−10	ACCA	ACCAAt	8,185
6. PTF2	−10	ACCA	ACCAT	8,038

Cluster motifs overlapping with known or putative TFBS −10 or −35 box motifs = 137 (total number of cluster motifs = 357).

R'MES stat = score of exceptionality (exceptionally frequent motifs will have high positive scores, whereas exceptionally rare motifs will have high negative scores).

PTF (1, 2, and 3): regulatory motif proposed by Studholme et al. (2004).

R = A/G; V = A/C/G; Y = C/T.

- A modified SigB consensus (ACGGTTT-N_{18}-TAC) was proposed. SigB is a general stress σ factor that has a binding site consensus ANGNNT-$N_{14–16}$-GGGTA (Cho et al., 2001).
- Two new consensuses for PTF2 were found. PTF2 is a regulatory motif proposed by Studholme et al. (2004) to be associated with the sigma factor regulation or expression.

3.4. Advantages of the new data mining method

This method is innovative in its use of the second-order Markov model to capture stochastic dependencies in k*mers* (see Section 2.1). HMM2 shows good performance in modeling the DNA heterogeneities and possesses three main advantages:

- It successfully detects small motifs (5–16 bases), minimizing the elusion of important regulatory sequences.
- It efficiently detects statistically exceptional motifs in the genome (e.g., the sigma factor BldN binding site).
- It proves to be versatile. This is why this method could be applied to the detection of DNA motifs other than SFBSs. Some of our detected motifs correspond to regulatory motifs such as those proposed by Studholme et al. (2004). Also, the ribosome binding sites (RBS), which are translational signals, have also been detected during this analysis. The *S. coelicolor* genome contains 2,576 RBSs annotated by Bentley et al. (2002)—2,041 of them in intergenic regions. Among them, 236 (12.9%) were revealed by our HMM2. Further, the same background HMM2 was used to detect various DNA repeats in the *S. coelicolor* genome (Hergalant et al., 2002).

TABLE 4. KNOWN TFBSs GENERATED BY THE DATA MINING ALGORITHM

	Function/involvement	*Consensus already defined*	*Consensus predicted*
R'MES score ≥ 6 $14 \leq spacer \leq 20$ $3 \leq l_{box2} \leq 5$			
WhiG	Early sporulation	**T**RVR-N_{14-16}-**SCCAG**NNW	GG**T**GAAT-N_{18}-**AG**T TGCGAA-N_{14}-CAG GCCGTAGG-N_{15}-GGCC
SigB	Osmotic stress response and aerial hyphae differentiation (sigma factor)	**ANG**NN**T**-N_{14-16}-**GGGTA**	**ACG**GT**T**T-N_{18}-**TA**C
SigR	Oxidative stress response (ECF sigma factor)	**GGAAT**-N_{18}-**GTT**	CG**GGAAT**-N_{18}-**GTT** TC**GGAAT**-N_{16}-TG**GT** GG**GGAAT**-N_{17}-G**GTT** CCC**GGAAT**-N_{17}-G**GT** AG**GGAAT**-N_{18}-**GTT**
R'MES score ≥ 4 $3 \leq spacer \leq 20$ $3 \leq l_{box2} \leq 4$			
PTF2 (Studholme)	Proposed to be associated with sigma factor regulation/expression	**TGAC**-N_{19}-**TGAC**	**TGAC**TTT-N_{16}-**TGA**

PTF2, regulatory motif proposed by Studholme et al. (2004).
N = A/C/G/T; S = G/C; R = A/G; V = A/C/G; W = A/T.

3.5. *Parameters influencing the behaviour of the HMM2*

3.5.1. Influence of the GC%. The genome of the *S. coelicolor* is GC rich (72.12%), and the intergenic regions present a slightly lower GC content (68.34%). In addition, the GC% of the extracted DNA motifs was 48.27%. Therefore, it may appear that the GC content may strongly influence our HMM2. To test this hypothesis, we classified the *iPeak-motifs* into 20 classes based on their GC%. Only one quarter of the *iPeak-motifs* showed a GC% in the 45–55 range, which includes the mean GC% (48.27%), and one third of them had a GC% scattered between the 40–45 and 55–60 range (for additional data, see distribution of *iPeaks-motifs* GC content; see online Supplementary Material at *www.liebertonline.com*). Furthermore, when the GC content of the intergenic sequence was observed along with the output probabilities, GC falls did not always generate peaks (data not shown). Thus, it can be concluded that the GC content does not have an exclusive influence on the model.

3.5.2. Influence of the pyrimidine/purine step and DNA repeats. DNA curvature plays a well-characterized role in many protein/DNA interactions and also more precisely in transcriptional regulation mechanisms (Jáuregui et al., 2003). It is also known that the flexibility of the DNA backbone is influenced by pyrimidine-purine steps (YR steps) (Bertrand et al., 1998). The percentage of YR steps determined within the intergenic regions of the genome (24.97%) and within the *iPeak-motifs* (23.58%) had a significant difference with a 95% confidence value. These results show that the YR step has an influence on the model (for additional data, see online Supplementary Material at *www.liebertonline.com*).

We also tested whether HMM2 was influenced by the presence of DNA repeats in the *iPeak-motifs*. Only 431 out of the 16,913 *iPeak-motifs* are found more than twice in the same intergenic region. Thus, HMM2 does not seem to be strongly influenced by local repeats of the same *iPeak-motif*. In contrast, we noticed that 8,520 *iPeak-motifs* (50.37%) were repeated at least once in another intergenic region. This result by itself is consistent and promising regarding the possibility of describing groups of co-regulated genes.

3.6. *Preliminary application of the method to the bacterial model* Bacillus subtilis

The above software suite was applied to the *B. subtilis* genome (Kunst et al., 1997). The main change to the model was in the HMM2+/− specification due to the GC skew of its genome (see Section 2.1). The GC% of

B. subtilis (roughly 43%) is distinct to one's of *S. coelicolor* (72%), but as demonstrated in section 3.2.1, the GC content does not exclusively influence the behavior of the model. We carried out a preliminary analysis of the *B. subtilis* genome using the same parameters (HMM2 topology, clustering parameters, R'MES score, motif structure) than for *S. coelicolor* analysis. The HMM2 model (four states, 3-*mors*, peak-variation set to 20%) generated 19,507 *iPeaks*. The *iPeak-motifs* were clustered into 360 clusters. Among them, 191 clusters had a R'MES score higher than six and were further considered as significant motifs. Each of them was successively considered as *box1* (-35 box and -10 box) to find the second *box2* (-10 box and -35 box, respectively) by the SFBS consensus search algorithm in order to produce *box1-spacer-box2* consensuses. The length of the spacer was set between 14 and 25 nt and the length of the *box2* (l_{box2}) was set between three and five. Finally, the data mining process generated 2,317 and 2,342 putative TFBS motifs respectively depending if we used the significant motifs as *box1* or *box2*. The genes that are controlled by the biologically described SigD, E, F, G, H, W, and X sigma factors (Makita et al., 2004) were retrieved with a ratio ranging from 25% to 52% (for additional data, see table of *B. subtilis* preliminary results; see online Supplementary Material at *www.liebertonline.com*).

5. CONCLUSION

We have proposed a hybrid data mining method based on stochastic and combinatorial algorithms to find SFBSs. HMM2 was processed on the whole genome without prior assumptions about its genetic organization. As already shown in speech recognition, data mining and ecology, the modeling of the hidden process by a second-order Markov chain exhibited a good capability in representing short segments such as SFBSs and other DNA motifs involved a transcriptional regulation. The parameters of the data mining process were tuned on a subpart of the genome—the *set-sco*—defined as a biologically well characterized regulon (SigR). On this subset, the HMM2 outperform the HMM1. On the actinomycete *S. coelicolor* genome, some already known promoters (SFBS) were found that strengthen the validity of the method, and we suggest a list of putative promoters (TFBS). Furthermore, the preliminary results from *B. subtilis* are promising with respect to the generalization of the method to other bacterial genomes.

AUTHORS' CONTRIBUTIONS

All authors read and approved the final manuscript. C.E. and S.H. adapted the stochastic data mining methods, which were initially developed by J.F.M. (the CarottAge toolbox), to the specificities of the genomic data and implemented the HMM2 related tools. C.A. carried out the different algorithms implementation and designed the TFBS search algorithm. B.A. and P.L. introduced the validation method, analysis, and interpretation of results and contributed useful knowledge of *Streptomyces* and other bacterial genetics.

ACKNOWLEDGMENTS

This project was supported by the ACI IMP-Bio initiative (Action Concertée Incitative, Informatique-Mathématiques-Physique en Biologie moléculaire, Ministère de l'Education Nationale, Ministère de l'Enseignement Supérieur et Ministère de la Recherche). This research was also supported by the INRA (Institut National de la Recherche Agronomique) and the Région Lorraine, France.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Ainsa, J.A., Parry, H.D., and Chater, K.F. 1999. A response regulator-like protein that functions at an intermediate stage of sporulation in *Streptomyces coelicolor* A3(2). *Mol. Microbiol*. 34, 607–619.

Bailey, T.L., and Elkan, C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. 2nd Int. Conf. Intellig. Syst. Mol. Biol.* 28–36.

Baum, L.E., Petrie, T., Soules, G., et al. 1970. A maximization technique occurring in statistical analysis of probabilistic functions in Markov chains. *Ann. Math. Statist*. 41, 164–171.

Bentley, S.D., Chater, K.F., Cerdeño-Tárraga, A.M., et al. 2002. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* 417, 141–147.

Bertrand, H.O., Ha-Duong, T., Fermandjian, S., et al. 1998. Flexibility of the B-DNA backbone: effects of local and neighbouring sequences on pyrimidine–purine steps. *Nucleic Acids Res*. 26, 1261–1267.

Besemer, J., Lomsadze, A., and Borodovsky, M. 2001. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res*. 29, 2607–2183.

Bibb, M.J., Molle, V., and Buttner, M. 2000. Sigma(BldN), an extracytoplasmic function RNA polymerase sigma factor required for aerial mycelium formation in *Streptomyces coelicolor* A3(2). *J. Bacteriol.* 182, 4606–4616.

Bize, L., Muri, F., Samson, F., et al. 1999. Searching gene transfers on *Bacillus subtilis* using Hidden Markov Models. *Proc. RECOMB 1999*, 43–49.

Blanchette, M., and Tompa, M. 2002. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res*. 12, 739–748.

Buhler, J., and Tompa, M. 2001. Finding motifs using random projections. *Proc. RECOMB 2001* 69–76.

Burden, S., Lin, Y.-X., and Zhang, R. 2005. Improving promoter prediction for the NNPP2.2 algorithm: a case study using *Escherichia coli* DNA sequences. *Bioinformatics* 21, 601–607.

Buttner, M.J., Smith, A.M., and Bibb, M.J. 1988. At least three different RNA polymerase holoenzymes direct transcription of the agarase gene (dagA) of *Streptomyces coelicolor* A3(2). *Cell* 52, 599–607.

Carvalho, A.M., Freitas, A.T., Oliveira, A.L., et al. 2004. Efficient extraction of structured motifs using box-links. *String Process. Inform. Retriev. Conf*. 267–278.

Chater, K.F., Bruton, C.J., Plaskitt, K.A., et al. 1989. The developmental fate of *S. coelicolor* hyphae depends upon a gene product homologous with the motility σ factor of *B. subtilis. Cell* 59, 133–143.

Cho, Y.H., Lee, E.J., Ahn, B.E., et al. 2001. SigB, an RNA polymerase sigma factor required for osmoprotection and proper differentiation of *Streptomyces coelicolor. Mol. Microbiol.* 42, 205–214.

Churchill, G. 1989. Stochastic models for heterogeneous DNA sequences. *Bull. Math. Biol*. 51, 79–94.

Corpet, F. 1988. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res*. 16, 10881–10890.

Dempster, A.P., Laird N.M., and Rubin D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc.* 39, 1–38.

Dsouza, M., Larsen, N., and Overbeek, R. 1997. Searching for patterns in genomic data. *Trends Genet*. 13, 597–498.

Du Preez, J.A. 1998. Efficient training of high-order hidden Markov model using first-order representations. *Comput. Speech Lang*. 12, 23–39.

Emmert, D.B., Stoehr, P.J., Stoesser, G., et al. 1994. The European Bioinformatics Institute (EBI) databases. *Nucleic Acids Res*. 26, 3445–3449.

Eskin, E., and Pevzner, P.A. 2002. Finding composite regulatory patterns in DNA sequences. *Bioinformatics* 13, 354–363.

Gordon, J.J., Towsey, M.W., Hogan, J.M., et al. 2006. Improved prediction of bacterial transcription start sites. *Bioinformatics* 22, 142–148.

He, Y. 1988. Extended Viterbi algorithm for second-order hidden Markov process. *Proc. IEEE Int. Conf. Pattern Recogn*. 2, 718–720.

Hergalant, S., Aigle, B., Decaris, B., et al. 2002. Intragenomic reiterations detection using Hidden Markov models. *Intellig. Syst. Mol. Biol*. 120.

Hiard, S., Maree, R., Colson, S., et al. 2007. PREDetector: a new tool to identify regulatory elements in bacterial genomes. *Biochem. Biophys. Res. Commun*. 357, 861–4.

Hoebeke, M., and Schbath, S. 2006. *R'MES: finding exceptional motifs. User guide, version 3*.

Hu, J., Li, B., and Kihara, D. 2005. Limitations and potentials of current motif discovery algorithms. *Nucleic Acids Res*. 33, 4899–4913.

Ikeda, H., Ishikawa, J., Hanamoto, A., et al. 2003 Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis. Nat. Biotechnol.* 21, 526–531.

Jacques, P.E., Rodrigue, S., Gaudreau, L., et al. 2006. Detection of prokaryotic promoters from the genomic distribution of hexanucleotide pairs. *BMC Bioinform*. 7, 423.

Jarmer, H., Larsen, T.S., Krogh, A., et al. 2001. Sigma A recognition sites in the *Bacillus subtilis* genome. *Microbiology* 147, 2417–2424.

Jáuregui, R., Abreu-Goodger, C., Moreno-Hagelsieb, G., et al. 2003. Conservation of DNA curvature signals in regulatory regions of prokaryotic genes. *Nucleic Acids Res*. 31, 6770–6777.

Kanhere, A., and Bansal, M. 2005. A novel method for prokaryotic promoter prediction based on DNA stability. *BMC Bioinform*. 6, 1.

Krogh, A., Brown, M., Mian, I., et al. 1994. Hidden Markov models in computational biology. *J. Mol. Biol.* 235, 1501–1531.

Kullback, S., and Leibler, R.A. 1951. On information and sufficiency. *Ann. Math. Statist*. 22, 79–86.

Kunst, F., Ogasawara, N., Moszer, I., et al. 1997. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* 390, 249–256.

Lawrence, C., Altschul, S., Boguski, M., et al. 1993 Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* 262, 208–214.

Le Ber, F., Benoît, M., Schott, C., et al. 2006. Studying crop sequences With CarottAge, a HMM-based data mining software. *Ecol. Model*. 191, 170–195.

Li, H., Rhodius, V., Gross, C., et al. 2002. Identification of the binding sites of regulatory proteins in bacterial genomes. *Proc. Natl. Acad. Sci. USA* 99, 11772–11777.

Liu, X., Brutlag, D.L., and Liu, J.S. 2001. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput*. 127–138.

Loots, G., Ovcharenko, I., Pachter, L., et al. 2002. rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res*. 12, 832–839.

Maetschke, S.R., Towsey, M.W., and Hogan, J.M. 2006. Bacterial promoter modelling and prediction for *E. coli* and *B. subtilis* with Beagle. *Proc. WISB–2006*, 43–49.

Makita, Y., Nakao, M., Ogasawara, N., et al. 2004. DBTBS: database of transcriptional regulation in *Bacillus subtilis* and its contribution to comparative genomics. *Nucleic Acids Res*. 32, D75–D77.

Mari, J.F., Haton, J.P., and Kriouile, A. 1997. Automatic word recognition based on second-order Hidden Markov Models. *IEEE Trans. Speech Audio Process.* 5, 22–25.

Mari, J.F., and Le Ber, F. 2006. Temporal and spatial data mining with second-order Hidden Markov Models. *Soft Comput*. 10, 406–414.

McCue, L., Thompson, W., Carmack, C., et al. 2001. Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res*. 29, 774–782.

McGuire, A.M., Hughes, J.D., and Church, G.M. 2000. Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Res*. 10, 744–757.

Munch, R., Hiller, K., Grote, A., et al. 2005. Virtual Footprint and PRODORIC: an integrative framework for regulon prediction in prokaryotes. *Bioinformatics* 21, 4187–4189.

Mwangi, M.M., and Siggia, E.D. 2003. Genome wide identification of regulatory motifs in *Bacillus subtilis. BMC Bioinform*. 16, 4–18.

Nicolas, P., Bize, L., Muri, F., et al. 2002. Mining *Bacillus subtilis* chromosome heterogeneities using hidden Markov models. *Nucleic Acids Res.* 30, 1418–1426.

Osada, R., Zaslavsky, E., and Singh, M. 2004. Comparative analysis of methods for representing and searching for transcription factor binding sites. *Bioinformatics* 20, 3516–3525.

Paget, M.S., Kang, J.G., Roe, J.H., et al. 1998. sigmaR, an RNA polymerase sigma factor that modulates expression of the thioredoxin system in response to oxidative stress in *Streptomyces coelicolor* A3(2). *EMBO J*. 17, 5776–5782.

Paget, M.S., Leibovitz, E., and Buttner, M.J. 1999. A putative two-component signal transduction system regulates sigmaE, asigma factor required for normal cell wall intergrity in *Streptomyces coelicolor* A3(2). *Mol. Microbiol*. 33, 97–107.

Paget, M.S., Molle, V., Cohen, G., et al. 2001 Defining the disulphide stress response in *Streptomyces coelicolor* A3(2): identification of the sigR regulon. *Mol Microbiol.* 42, 1007–1020.

Petersen, L., Larsen, T.S., Ussery, D.W., et al. 2003. RpoD promoters in *Campylobacter jejuni* exhibit a strong periodic signal instead of a −35 box. *J. Mol. Biol*. 326, 1361–1372.

Potuckova, L., Kelemen, G.H., Findlay, K.C., et al. 1995. A new RNA polymerase sigma factor, sigma F, is required for the late stages of morphological differentiation in *Streptomyces* spp. *Mol. Microbiol*. 17, 37–48.

Rajewsky, N., Socci, N.D., Zapotocky, M., et al. 2002. The evolution of DNA regulatory regions for proteo-gamma bacteria by interspecies comparisons. *Genome Res*. 12, 298–308.

Rangannan, V., and Bansal, M. 2007. Identification and annotation of promoter regions in microbial genome sequences on the basis of DNA stability. *J. Biosci.* 32, 851–862.

Robison, K., McGuire, A.M., and Church, G.M. 1998. A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome. *J. Mol. Biol*. 284, 241–254.

Robin, S., Daudin, J.-J., Richard, H., et al. 2002. Occurrence probability of structured motifs in random sequences. *J. Comput. Biol.* 9, 761–773.

Rutherford, K., Parkhill, J., Crook, J., et al. 2000. Artemis: sequence visualisation and annotation. *Bioinformatics* 16, 944–945.

Ryding, N.J., Kelemen, G.H., Whatling, C.A., et al. 1998. A developmentally regulated gene encoding a repressor-like protein is essential for sporulation in *Streptomyces coelicolor* A3(2). *Mol. Microbiol*. 29, 343–357.

Segal, E., and Sharan, R. 2005. A discriminative model for identifying spatial cis-regulatory modules. *J. Comput. Biol*. 12, 822–834.

Servin-Gonzalez, L., Jensen, M.R., White, J., et al. 1994. Transcriptional regulation of the four promoters of the agarase gene (dagA) of *Streptomyces coelicolor* A3(2). *Microbiology* 140, 2555–2565.

Siddharthan, R., Siggia, E., and Van Nimwegen, E. 2005. PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput. Biol*. 1, e67.

Smith, T.F., and Waterman, M.S. 1981. Identification of common molecular subsequences. *J. Mol. Biol*. 147, 195–197.

Stormo, G.D. 2000. DNA binding sites: representation and discovery. *Bioinformatics* 16, 16–23.

Studholme, D.J., Bentley, S.D., and Kormanec, J. 2004. Bioinformatic identification of novel regulatory DNA sequence motifs in *Streptomyces coelicolor. BMC Microbiol*. 4, 14.

Thijs, G., Lescot, M., Marchal, K., et al. 2001. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics* 17, 1113–1122.

Tompa, M., Li, N., Bailey, T.L., et al. 2005. Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol*. 23, 137–144.

Touzain, F., Schbath, S., Debled-Rennesson, I., et al. 2008. A tool to search for sigma factor binding sites in bacterial genomes using comparative approach and biologically driven statistics. *BMC Bioinform*. 9, 73.

Typas, A., and Hengge, R. 2005. Differential ability of σS and σ70 of *Escherichia coli* to utilize promoters containing half or full UP-element sites. *Mol. Microbiol*. 55, 250–260.

Van Helden, J., Del Olmo, M., and Pérez-Ortín, J.E. 2000. Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. *Nucleic Acids Res*. 28, 1000–1010.

Vanet, A., Marsan, L., Labigne, A., et al. 2000. Inferring regulatory elements from a whole genome. An analysis of *Helicobacter pylori* sigma(80) family of promoter signals. *J. Mol. Biol*. 297, 335–53.

Wang, T., and Stormo, G.D. 2003. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics* 19, 2369–80.

Ward, J.H. 1963. Hierarchical grouping to optimize an objective function. *J. Am. Statist. Assoc*. 58, 236–244.

Yada, T., Totoki, Y., Ishikawa, M., et al. 1998. Automatic extraction of motifs represented in the hidden Markov model from a number of DNA sequences. *Bioinformatics* 14, 317–325.

Address correspondence to:
Prof. Pierre Leblond
Genetics and Microbiology Laboratory
INRA and Nancy University
Boulevard des aiguillettes
54500 Vandoeuvre-les-Nancy, France

E-mail: leblond@nancy.inra.fr

or

Prof. Jean-François Mari
LORIA
CNRS 7503 and INRA Grand Est
Boulevard des aiguillettes
54500 Vandoeuvre-les-Nancy, France

E-mail: jean-francois.mari@loria.fr